Bayesian Semiparametric Hierarchical Empirical Likelihood Spatial Models

Aaron T. Porter
Department of Statistics
University of Missouri

Joint Work With
Scott H. Holan, Christopher K. Wikle

NCRN Webinar
May 7, 2014

Outline

- Motivation for the Semiparametric Hierarchical Empirical Likelihood (SHEL) framework.
- Details of the framework.
- An introduction to Bayesian spatial models.
- Predicting socioeconomic indicators for Missouri counties from survey data.
- Conclusions

SHEL Motivation

- It is not uncommon in statistics to encounter datasets that do not follow a common parametric form.
- This may be due to a variety of issues with data (i.e. outliers, heavy or light tails, abnormal skewness).
- Additionally, mixtures of distributions may be present, but the underlying parametric forms may be unknown.
- Transformations may be difficult to identify in order to model such data.

SHEL Motivation

- We desire methodology to handle "poorly behaved" datasets that is flexible enough to handle spatial and other general dependence structures.
- The methodology should be general enough to handle data that are either continuous or discrete, and that exist on either a continuous or discrete support.
- We illustrate our model on three datasets:
  - data concerning per capita income in Missouri counties from the American Community Survey (ACS) are collected on an irregular areal support and represents a continuous outcome,
  - data from the North American Breeding Birds Survey which are collected over a continuous support and represents a count data, and
  - Sudden Infant Death Syndrome (SIDS) data which are collected on an irregular areal support and represent a count data.

Bayesian Hierarchical Models

- Let $\mathbf{Z}$ be a n_z-dimensional vector of observations, $\mathbf{Y}$ be an n_y-dimensional vector corresponding to an unobserved process, and $\boldsymbol{\xi} = \{\boldsymbol{\xi_D}, \boldsymbol{\xi_P}\}$ be a set of parameters related to both the data model and process model at hand.
- Further, let $[\mathbf{Z}|\mathbf{Y}]$ denote the conditional distribution of $\mathbf{Z}$ given $\mathbf{Y}$ and $[\mathbf{Y}]$ denote the marginal distribution of $\mathbf{Y}$. The basic Bayesian Hierarchical Model (BHM) can be written as:

$$\begin{aligned} \text{Data Model:} &[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi_D}] \\ \text{Process Model:} &[\mathbf{Y}|\boldsymbol{\xi_P}] \\ \text{Parameter Model:} &[\boldsymbol{\xi_D}, \boldsymbol{\xi_P}], \end{aligned}$$

Empirical Likelihood

- The original empirical likelihood (EL) of Owen (1988) serves to bypass the issue of selecting a parametric form for the data by broadly placing estimation of fixed effect parameters in a nonparametric context.
- We consider using the empirical cumulative distribution function (ECDF), where the ECDF is defined as $F_n(z) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{Z_i \leq z\}}$, for $Z_1, \ldots, Z_n \in \mathbb{R}$.
- If we assume $Z_1, \ldots, Z_n$ are independent and share a common CDF F, then the EL can be defined as $L(F_n) = \prod_{i=1}^{n}\{F(z_i) - F(z_{i-})\}$.
- What remains is to estimate the weights $w_i = \{F(z_i) - F(z_{i-})\}$.
- This formulation clearly avoids parametric specifications in favor of allowing the data to define the likelihood.

Empirical Likelihood

- There are many techniques for estimating the weights $\{w_i\}$, but we consider the estimating equations approach of Qin and Lawless (1994).
- The EL of a vector of functionals $\boldsymbol{\theta} = \{\theta_1, \ldots, \theta_l\}$ of the form $k_u(\sum_{i=1}^n w_i z_i) = \theta_u$, $u = 1, \ldots, l$, for a known function $k_u(\cdot)$ given iid $\{Z_i\}$ can be computed as:

$$L(\theta) = \prod_{i=1}^{n} w_i(\boldsymbol{\theta})$$

where $L(\theta)$ is maximized over the simplex

$$W_{\boldsymbol{\theta}} = \left\{ \sum_{i=1}^{n} w_i = 1; w_i > 0 \text{ for all } i \; ; \sum_{i=1}^{n} w_i m_j(z_i, \theta_i) = 0 \text{ for all } j \right\}.$$

Here, for i in $1, \ldots, n$, $[\{m_j(z_i, \theta_i)\} \in \mathbb{R}]$ for $j = 1, ..., J$ are a set of J estimating equations.

Empirical Likelihood

- Chaudhuri and Ghosh (2011) suggest the use of the estimating equations

$$\sum_{i=1}^{n} w_i\{z_i - \theta_i\} = 0$$

$$\sum_{i=1}^{n} \{w_i(z_i - \theta_i)^2 / V(\theta_i)\} - 1 = 0,$$

- Here θ_i represents $E(Z_i)$ and $V(\theta_i)$ represents $\text{Var}(Z_i|\theta_i)$.
- These estimating equations are based on the the exponential family of distributions.
- Thus far, we have assumed iid observations. In spatial applications, independence is an unrealistic assumption.

Empirical Likelihood

- Spatial EL literature is rather underdeveloped in comparison to other areas of EL.
- While the EL has been extended to handle certain types of dependence, the types of dependence present in spatial literature are generally difficult to accommodate in an observation-driven EL framework.
- The SHEL model we propose fills this gap by placing the EL in the BHM framework.
- This also serves to simultaneously solve the issue of the selection of a parametric data model in the BHM framework.

Semiparametric Hierarchical Empirical Likelihood

- We utilize the BHM framework to construct the SHEL model, which has the form:

$$\begin{aligned} \text{Empirical Data Model:} & [\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi_D}] \\ \text{Process Model:} & [\mathbf{Y}|\boldsymbol{\xi_P}] \\ \text{Parameter Model:} & [\boldsymbol{\xi_D}, \boldsymbol{\xi_P}]. \end{aligned}$$

- We further assume $E(\mathbf{Z}|\mathbf{Y}) = g(\mathbf{X}\boldsymbol{\beta} + \mathbf{Y})$ and $E(\mathbf{Z}^2|\mathbf{Y}) = h(\mathbf{X}\boldsymbol{\beta} + \mathbf{Y})$ for g and h known and $\mathbf{X}$ is an $n \times p$ design matrix of fixed and known covariate information.
- These relationships will serve to inform a set of estimating equations utilized in estimating the parameters involved in the empirical data model and the process model.

Semiparametric Hierarchical Empirical Likelihood

- As we alluded to, we account for the data dependency in the parametric process model, i.e., we assume that $[Z_i, Z_j|\mathbf{Y}] = [Z_i|\mathbf{Y}][Z_j|\mathbf{Y}]$ for all $i \neq j$.
- This allows us to use standard EL techniques, EL techniques at the data model stage with the independence assumption replaced with one of conditional independence.
- In the spatial modeling framework, we are able to alleviate the strict structure imposed by the restrictive blocking arguments typically used for spatial and temporal data.

Semiparametric Hierarchical Empirical Likelihood

- There are methodological and computational concerns that come with utilizing a nonparametric data model.
- One important aspect to consider is that the joint distribution $[\mathbf{Y}, \boldsymbol{\xi}]$ must be proper, as the non-analytic form of $[\mathbf{Z}|\mathbf{Y}, \boldsymbol{\xi}]$ makes verification of propriety difficult when $[\mathbf{Y}, \boldsymbol{\xi}]$ is improper.
  - This is a major concern in lattice data, where improper spatial priors are frequently used.
  - In our manuscript, we utilize a rank-reduced version of the intrinsic conditional autoregressive (ICAR) prior from Hughes and Haran (2013) and develop mathematical conditions to ensure propriety.
- There are additionally considerable computational challenges to be overcome, and these are discussed in the manuscript.

ICAR models and propriety conditions

- While the ICAR model provides the strongest spatial association of the commonly used Bayesian spatial models for areal data, it is degenerate.
- The ICAR can be defined as:

$$Y_i | Y_{j,j\neq i} \sim N\left(\sum_{j\in ne(i)}\left\{\frac{b_{ij}}{\sum_{j\in ne(i)} b_{ij}} y_j\right\}, \frac{1}{\tau \sum_{j\in ne(i)} b_{ij}}\right),$$

where $b_{ij} = 1$ if locations i and j are neighbors and 0 otherwise, and $j \in ne(i)$ indicates that locations i and j are neighbors. This argument yields a probability density function for $\mathbf{Y} = (Y_1, \ldots, Y_n)'$

$$\pi(\mathbf{Y} = \mathbf{y}) \propto \exp\left\{-\frac{1}{2}\mathbf{y}'\tau(\mathbf{B}_+ - \mathbf{B})\mathbf{y}\right\},$$

where $\mathbf{B}_{ij} = b_{ij}$ and $\mathbf{B}_+$ is a diagonal matrix with $\mathbf{B}_{+,ii} = \sum_{j=1}^{n} b_{ij}$.

ICAR models and propriety conditions

- The "precision" matrix $(\mathbf{B}_+ - \mathbf{B})$ is subject to the constraint $(\mathbf{B}_+ - \mathbf{B})\mathbf{1} = \mathbf{0}$ and is of rank $n-1$, and therefore not invertible.
- There are several possible solutions to this problem, but our preferred starting point is the rank reduction found in Hughes and Haran (2013).
- They consider a process $\mathbf{Y}_n = \mathbf{M}_{n\times q}\mathbf{Y}^*_q$, where $\pi(\mathbf{Y}^* = \mathbf{y}^*) \propto \tau^{\frac{q}{2}} \exp\{-\frac{1}{2}\tau \mathbf{y}^{*\prime}\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}\mathbf{y}^*\}$ with $\mathbf{M}$ being a matrix with columns equal to the eigenvectors of $(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{B}(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')$ associated with the largest q eigenvalues of the matrix.
- The work provided a rank-reduced solution to a well-known confounding issue in ICAR modeling, and also provides a starting point for building a proper lattice prior.

ICAR models and propriety conditions

- **Theorem**: Consider a hierarchical Bayesian model in which the data model has two finite moments $E(\mathbf{Z}|\mathbf{Y}) = g(\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{Y}^*)$ and $E(\mathbf{Z}^2|\mathbf{Y}) = h(\mathbf{X}\boldsymbol{\beta} + \mathbf{M}\mathbf{Y}^*)$ for g and h known. Let the process $\mathbf{Y}^*$ be given a Hughes and Haran (2013) prior of the form $\pi(\mathbf{Y}^* = \mathbf{y}^*) \propto \tau^{\frac{q}{2}} \exp\{-\frac{1}{2}\tau \mathbf{y}^{*'}\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}\mathbf{y}^*\}$, where $\text{rank}(\mathbf{M}) \leq n - 1$. Assume that $\mathbf{B}$ is the adjacency matrix for a first order ICAR (i.e., $\text{rank}(\mathbf{B}) = n - 1$). Then, a sufficient condition for $\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}$ to be positive definite is that the design matrix $\mathbf{X}$ contains the one vector.
- The proof of this theorem can be found in our manuscript.
- This condition is sufficiently weak to be utilized for most applications of spatial regression models for areal data.

Small Area Estimation

- The dataset that we primarily work with is the American Community Survey (ACS), which is an ongoing survey conducted by the U.S. Census Bureau.
- The ACS provides period estimates of the variables on the long form of the decennial census, as well as other variables. These estimates may be in the form of one year, three year, or five year period estimates, depending on the population of the area.
- Small area estimation (SAE) is a set of statistical methods to improve the precision of undersampled (unsampled) geographies.
- For undersampled (unsampled), such as smaller geographies, SAE is needed.
- Sometimes, state-level data can be highly variable and SAE methods can still provide improved estimates over the survey itself.

The FH model

- For our particular data analysis, we will work at the county level in the state of Missouri to perform SAE.
- One of the key tools for small area estimation is the Fay-Herriot (FH) model, due to Fay and Herriot (1979).

$$Z_i = \theta_i + \epsilon_i$$
$$\theta_i = \mathbf{x}_i'\boldsymbol{\beta} + y_i$$

- Here, Z_i is a design-unbiased survey estimate of the superpopulation parameter of interest, θ_i, at location i, $\mathbf{x}_i$ is auxiliary information, y_i is a random effect associated with location i, and ϵ_i is the sampling error at location i with known variance σ_i^2.

Sampling Errors

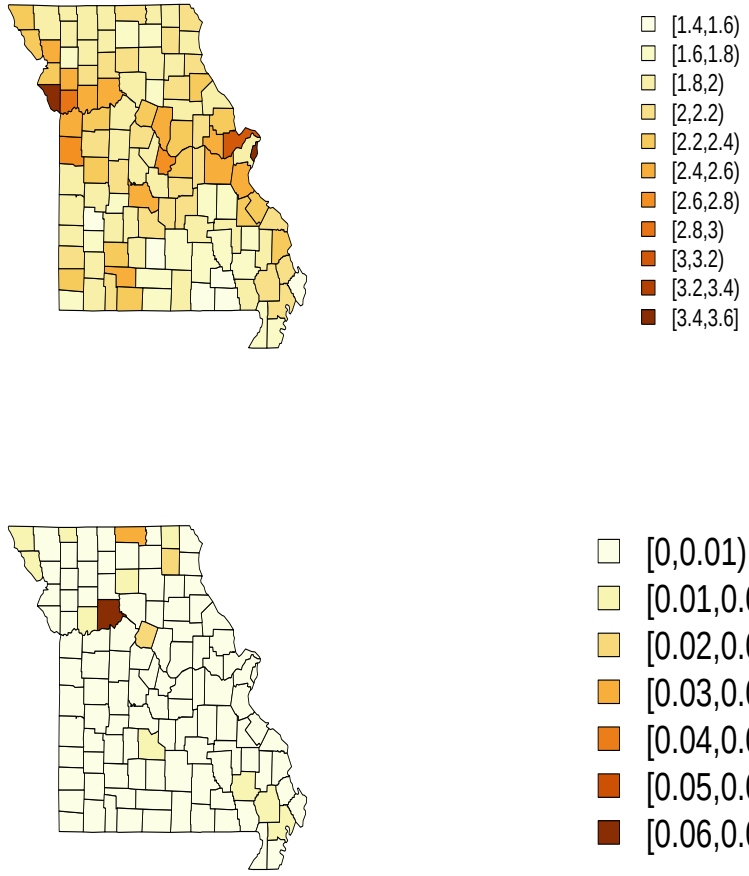
- Traditionally, one assumes $y_i \sim_{iid} N(0, \sigma^2)$, though spatial models have become more popular in the past five years.
- Typically, one also assumes $\{\epsilon_i\}$ are independent, normally distributed errors with mean 0 and variance σ_i^2.
- Chaudhuri and Ghosh (2011) deviate from this assumption on $\{\epsilon_i\}$ in the small area framework by replacing the normality assumption with an EL model in a hierarchical nested error regression model.
- They additionally utilize $y_i \sim_{iid} N(0, \sigma^2)$ for one analysis and $y_i|G \sim G,\ G|A \sim \text{DP}(\alpha, \mathcal{G})$, where DP represents a Dirichlet Process prior for a second analysis and $\mathcal{G}$ is a Gaussian base measure.
- This model can be seen as a special case of our SHEL framework.

The ACS Data

- Our goal is to estimate mean per capita income for Missouri counties.
- The data comes in the form of a five-year period estimate from 2006-2010.
- For auxiliary information, we use the five-year period estimate of percent unemployed in each county.
- The weighted least squares (WLS) residuals cannot be transformed to Gaussian with a Box-Cox or a log transformation.
- The residuals do demonstrate spatial autocorrelation.

The ACS Data

Figure: Mean per capita income in all 115 Missouri counties and the sampling variance of each estimate.



The ACS Data

- In analyzing these data, we consider five possible specifications:
  - Our SHEL specification with a Hughes and Haran (2013) prior.
  - The EL nested error regression model of Chaudhuri and Ghosh (2011) with an independence prior on $\mathbf{Y}$.
  - The EL nested error regression model of Chaudhuri and Ghosh (2011) with a DP prior on $\mathbf{Y}$.
  - A standard FH analysis with a Gaussian likelihood.
  - A spatial (Hughes and Haran, 2013) FH analysis with a Gaussian likelihood.
- We use vague prior for all analyses, with the exception of the DP prior model, which required an informative prior on the Gaussian base measure variance.

The ACS Data

- The actual prior specifications utilized were:
  - SHEL: $\mathbf{Y} \sim N(0, \tau^{-1}\{\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}\}^{-1})$; $\beta \sim \mathrm{N}(\beta^*, g^{-1}\tau^{-1}I_2)$; $\tau \sim$ Gamma(1,1).
  - Ind EL: $Y_i \sim_{iid} N(0, A)$; $\beta \sim \mathrm{N}(\beta^*, g^{-1}AI_2)$; $A \sim$ IG(1,1).
  - DP EL: $Y_i|G \sim G$, $G|A \sim \mathrm{DP}(\alpha, \mathcal{G})$; $\beta \sim \mathrm{N}(\beta^*, g^{-1}AI_2)$; $A \sim$ IG(2,1000).
  - Independence FH: $Y_i \sim_{iid} N(0, A)$; $\beta \sim \mathrm{N}(\beta^*, g^{-1}AI_2)$; $A \sim$ IG(1,1).
  - Spatial FH: $\mathbf{Y} \sim N(0, \tau^{-1}\{\mathbf{M}'(\mathbf{B}_+ - \mathbf{B})\mathbf{M}\}^{-1})$; $\beta \sim \mathrm{N}(\beta^*, g^{-1}\tau^{-1}I_2)$; $\tau \sim$ Gamma(1,1).
- Where g is Zellner's g-prior, specified as in Chaudhuri and Ghosh (2011), and β^* to be the WLS estimate of β.

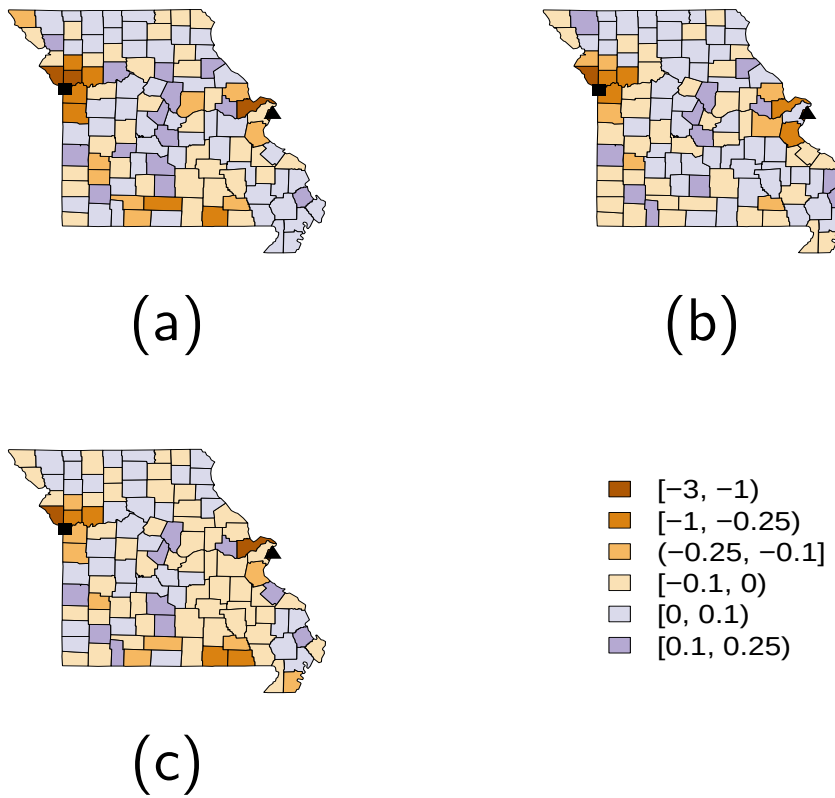
The ACS Data

Model	β_0	β_1	A	MSPE
SHEL	2.164 (2.051, 2.256)	-0.042 (-0.063, -0.015)	0.287 (0.157, 0.628)	0.066
Independence EL	2.230 (2.210, 2.364)	-0.077 (-0.095, -0.058)	0.008 (0.004, 0.015)	0.128
DP EL	2.331 (2.170, 2.474)	-0.0375 (-0.069, -0.002)	0.049 (0.006, 0.745)	0.128
Independence Parametric	2.094 (1.971, 2.217)	-0.006 (-0.027, 0.015)	0.142 (0.109 0.187)	0.130
Spatial Parametric	2.327 (2.284, 2.370)	-0.058 (-0.067, -0.050)	0.503 (0.345, 0.765)	0.076

Table: Posterior medians and 95% (central) credible intervals for the FH example (Section 5.1). A represents the variance of $\mathbf{y}$ in the Chaudhuri and Ghosh (2011) parameterizations, and τ^{-1} for the SHEL parameterization. MPV is the mean posterior variance of $\boldsymbol{\theta}$ for each model.

The ACS Data

Figure: The difference of the squared deviations $(Y_i - \widehat{Y}_{(-i)})^2$ for each location of estimated per capita income for (a) the SHEL model versus the Chaudhuri and Ghosh (2011) independence model, (b) the SHEL model versus the Chaudhuri and Ghosh (2011)DP model, (c) the SHEL model versus the parametric model. The square represents Kansas City, MO and the triangle represents St. Louis, MO.



The ACS Data

- The SHEL model provides much better model fit in terms of leave-one-out MSPE than other model choices, providing a 48% reduction in MSPE over the best fitting independence model.
- The SHEL model provides a 13% reduction in MSPE over the parametric FH model explicitly accounting for spatial correlation.
- Explicitly accounting for spatial correlation is key for these data.

Conclusions

- The SHEL model overcomes the difficulty of selecting a data distribution in the traditional BHM framework.
- The SHEL model provides a natural way to incorporate spatial and other dependence structures in the EL without the need for restrictive blocking arguments.
- The SHEL model can be used for continuous or discrete outcomes over continuous or discrete spatial and temporal supports.
- Simulations and data analyses illustrate the SHEL models as providing strong predictive performance relative to parametric models.
- The SHEL model can handle messy, correlated data in a natural way, with fewer assumptions that risk being violated.

Thank You!

porterat@missouri.edu

Relevant References:

Chaudhuri, S. and Ghosh, M. (2011). Empirical likelihood for small area estimation. Biometrika, 98(2):473-480.

Fay, R. and Herriot, R. (1979). Estimates of income for small places: an application of James-Stein procedures to census data." Journal of the American Statistical Association, 74, 269-277.

Hughes, J. and Haran, M. (2013). Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 75(1):139-159.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. Biometrika, 75(2):237-249.

Porter, A. T., Holan, S. H., and Wikle, C. K. "Bayesian semiparametric hierarchical empirical likelihood spatial models." Submitted.